# VQEG SUBJECTIVE TEST PLAN

Prepared by Philip Corriveau & Nikolaus Walch (Communications Research Centre)
and Alexander Schertz (IRT)

# 1 INTRODUCTION

A group of experts from three groups, ITU-R SG11, ITU-T SG9 and ITU-T SG12 assembled in Turin Italy on 14-16 October 1997 to form the Video Quality Experts Group (VQEG). The goal of the meeting was to create a framework for the evaluation of new objective methods for video quality evaluation. Four groups were formed under the VQEG umbrella: Independent Labs and Selection Committee (ILSC), Classes and Definitions, Objective Test Plan, and Subjective Test Plan. In order to assess the correlations between objective and subjective methods, a detailed subjective test plan has been drafted.

A second meeting of the video Quality Experts Group took place in Gaithersburg USA on 26-29 May 1998 at which time a first draft of the subjective test plan was finalised.

The purpose to subjective testing is to provide data on the quality of video sequences and to compare the results to the output of proposed objective measurement methods. This test plan provides common criteria and a process to ensure valid results from all participating facilities.

# 2 THE DOUBLE-STIMULUS CONTINUOUS QUALITY-SCALE METHOD

The Double Stimulus Continuous Quality Scale (DSCQS) method will be used because it's the most reliable (with respect to contextual effects) and most widely used procedure proposed by Rec. ITU-R BT.500-8.

## 2.1 GENERAL DESCRIPTION

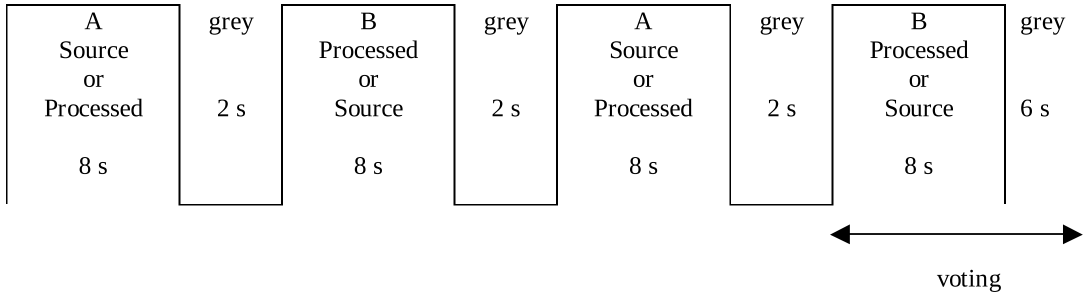


FIGURE 1 PRESENTATION STRUCTURE OF TEST MATERIAL

The DSCQS method presents two pictures (twice each) to the assessor, where one is a source sequence and the other is a processed sequence (see Figure 1). A source sequence is unimpaired whereas a processed sequence may or may not be impaired. The sequence presentations are randomised on the test tape to avoid the clustering of the same conditions or sequences. Participants evaluate the picture quality of both sequences using a grading scale (DSCQS). They are invited to vote as the second presentation of the second picture begins and are asked to complete the voting before completion of the grey period after that.

## 2.2 GRADING SCALE

The DSCQS consists of two identical 10 cm graphical scales which are divided into five equal intervals with the following adjectives from top to bottom: Excellent, Good, Fair, Poor and Bad. (Note: adjectives will be written in the language of the country performing the tests.) The scales are positioned in pairs to facilitate the assessment of each sequence, i.e. both the source and processed sequence. The viewer records his/her assessment of the overall picture quality with the use of pen and paper or an electronic device (e.g. a pair of sliders). Figure 4, shown below, illustrates the DSCQS.

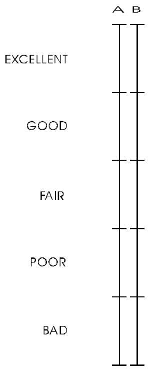


FIGURE 2 DSCQS (NOT TO SCALE)

## 3 TEST MATERIALS

### 3.1 SELECTION OF TEST MATERIAL

Twenty sequences were selected by the ILSC to be used in the test: half of them are in 50 Hz format and half in 60 Hz format. In addition, two sequences that are available both in 50 and 60 Hz formats were selected for the training. The sequences were provided by RAI, CCETT, AT&T and CRC. The factors[1] taken into account in the selection were:

a) Colour
- at least one sequence must stress the colour
- saturated colours should be on moving and 'important' objects, that are objects that attract the attention of the viewer
- different skin colours
- colour patterns moving around while luminance is not changing (this characteristics is not realised in any of the test scenes)

b) Luminance
- high luminance
- low luminance

c) several film sequences

d) several sequences containing motion energy and spatial detail:
- at least one still picture should be included. Concerning still pictures it was agreed that at least one of the following characteristics should be represented:
  - – different directions
  - – saturated colours
  - – maybe also text (english) if critical fonts are used
- zooming
- an object that appears and crosses quickly the scene
- areas moving in different directions (e.g. camera panning + an object that is moving in all directions
- text scrolling both in vertical and Horizontal direction
- in general only one scene cut will be accepted, but a few sequences (2-3) should have several scene cuts
- All the directions should be represented

e) Source
- ITU-R BT.601
- down-converted
- analogue component

[1] It was not required that every factor is represented both in 50 and 60 Hz set of sequences

- film
- synthetic

f) General

- Test conditions must span the whole quality range
- Scene content will either facilitate or mask certain forms of degradation when present (e.g. flat areas, complex patterns, square patterns, water motion, broad range of sequences)
- Culturally neutral and gender 'unbiased'

Table 1 lists the chosen sequences.

The use of the VQEG test sequences shall be restricted to the VQEG evaluation technical tests and shall not be re-used without permission for any other other purpose and in any other form, including the development, promotion, demonstration and commercialisation of products directly or indirectly derived from the VQEG activities. It shall not be used without permission for any non-VQEG related evaluations,developments and /or commercial purposes (including demonstration and promotion).

TABLE 1: List of selected sequences

625/50 format

| Assigned number | Sequence | Characteristics | Source |
|---|---|---|---|
| 1 | Tree | Still, different direction | EBU |
| 2 | Barcelona | Saturated colour + masking effect | RAI & Retevision [2] |
| 3 | Harp | Saturated colour, zooming, highlight, thin details | CCETT |
| 4 | Moving graphic | Critical for Betacam, colour, moving text, thin characters, synthetic | RAI |
| 5 | Canoa Valsesia | water movement, movement in different direction, high details | RAI |
| 6 | F1 Car | Fast movement, saturated colours | RAI |
| 7 | Fries | Film, skin colours, fast panning | FILM [3] |
| 8 | Horizontal scrolling 2 | text scrolling | RAI |
| 9 | Rugby | movement and colours | RAI |
| 10 | Mobile&calendar | available in both formats, colour, movement | CCETT |
| 11 | Table Tennis | Table Tennis (training) | CCETT |
| 12 | Flower garden | Flower garden (training) | CCETT/KDD |

525/60 format

| Assigned number | Sequence | Characteristics | Source |
|---|---|---|---|
| 13 | Baloon-pops | film, saturated colour, movement | FILM[4] |
| 14 | NewYork 2 | masking effect, movement) | AT&T[5] |
| 15 | Mobile&Calendar | available in both formats, colour, movement | CCETT |
| 16 | Betes_pas_betes | colour, synthetic, movement, scene cut | |
| 17 | Le_point | colour, transparency, movement in all the directions | |
| 18 | Autumn_leaves | colour, landscape, zooming, water fall movement | |
| 19 | Football | colour, movement | |
| 20 | Sailboat | almost still | EBU? |
| 21 | Susie | skin colour | EBU? |
| 22 | Tempete | colour, movement | EBU? |
| 23 | Table Tennis (training) | Table Tennis (training) | CCETT |
| 24 | Flower garden (training) | Flower garden (training) | CCETT/KDD |

[2] The sequence was produced by RAI in collaboration with Spanish TV
[3] Provided by RAI
[4] Provided by CCETT
[5] Provided by CSELT

## 3.2 HYPOTHETICAL REFERENCE CIRCUITS (HRC)

TABLE 2: HRC list

| ASSIGNED NUMBER | A | B | BIT RATE | RES | METHOD | COMMENTS |
|---|---|---|---|---|---|---|
| 16 | X | | 1.5 Mb/s | CIF | H.263 | Full Screen |
| 15 | X | | 768 kb/s | CIF | H.263 | Full Screen |
| 14 | X | | 2 Mb/s | ¾ | mp@ml | This is horizontal resolution reduction only |
| 13 | X | | 2 Mb/s | ¾ | sp@ml | |
| 12 | X | | 4.5 Mb/s | | mp@ml | With errors TBD |
| 11 | X | | 3 Mbit/s | | mp@ml | With errors TBD |
| 10 | X | | 4.5 Mb/s | | mp@ml | |
| 9 | X | X | 3 Mbit/s | | mp@ml | |
| 8 | X | X | 4.5 Mb/s | | mp@ml | Composite NTSC and/or PAL |
| 7 | | X | 6 Mb/s | | mp@ml | |
| 6 | | X | 8 Mb/s | | mp@ml | Composite NTSC and/or PAL |
| 5 | | X | 8 & 4.5 Mb/s | | mp@ml | Two codecs concatenated |
| 4 | | X | 19/PAL(NTSC)-19/PAL(NTSC)-12 Mbit/s | | 422p@ml | PAL or NTSC 3 generations |
| 3 | | X | 50-50-…<br>-50 Mbits/s | | 422p@ml | 7th generation with shift / I frame |
| 2 | | X | 19-19-12 Mbit/s | | 422p@ml | 3rd generations |
| 1 | | X | n/a | | n/a | Multi-generation Betacam with drop-out(4 or 5, composite/component) |

Details of the HRCs are given in ANNEX 2.

## 3.3 SEGMENTATION OF TEST MATERIAL

Since there are two standard formats 525/60 and 625/50, the test material could be split 50/50 between them.

The range of quality that is to be examined in this test is extremely large and not done conventually with one test. In order to avoid having compressed quality judgments in the High Quality range it was decided to have two separate tests, one with High Quality processed sequences and one with Low Quality processed sequences. As the DSCQS method (see chapter 2) involves the assessment of processed and reference sequences, the test with Low Quality processed sequences will include High Quality sequences as well (namely the references). So we will have a broad range quality test (called "Low Quality test" in the following) and a narrow range quality test (called "High Quality test" in the following). It is possible that there might be one or more models suited to the narrow High Quality region but not to the broad region involved in the Low Quality test. This design is not perfect but is a compromise to help achieve most of the goals of this test. Details on the discussion inside VQEG on this problem are given in ANNEX 4.

Therefore, the first test will be done using a low bit rate range of 768 kb/s – 4.5 Mb/s (16,15,14,13,12,11,10,9,8) Table 1 for a total of 9 HRC's. A second test will be done using a high bit rate range of 3 Mb/s -50 Mb/s (9,8,7,6,5,4,3,2,1) Table 1 for a total of 9 HRC's. It can be noted that 2 conditions (9 & 8) are common to both test sets.

### 3.3.1 DISTRIBUTION OF TESTS OVER FACILITIES

There was a long discussion whether 50 Hz tests should be restricted to 50 Hz countries or not. Several members of VQEG were concerned that people accustomed to 525/60 television may perceive the flicker of 625/50 pictures more easily than people from 50 Hz countries and this may bias the results. Because it turned out that there was a shortage of labs in 60 Hz countries whereas enough labs in 50 Hz countries were ready to participate, this discussion became irrelevant. Agreement was reached on the following distribution:

| Laboratory Code | TEST SITE | 50Hz tests | 60Hz tests |
|---|---|---|---|
| 1 | Berkom (FRG) | | X |
| 2 | CRC (CAN) | | X |
| 3 | FUB (IT) | | X |
| 4 | NHK (JPN) | | X |
| 5 | CCETT (FR) | X | |
| 6 | CSELT (IT) | X | |
| 7 | DCITA (AUS) | X | |
| 8 | RAI (IT) | X | |

Assuming that in any laboratory at least 15 subjects will participate in the tests, as a result of this distribution of work there will be a total of 60 subjects running 50 Hz and other 60 subjects running 60 Hz tests.

Each test tape will be assigned a number so that we are able to track which facility conducts which test.. The tape number will be inserted directly into the data file so that the data is linked to one test tape.

### 3.3.2 PROCESSING AND EDITING SEQUENCES

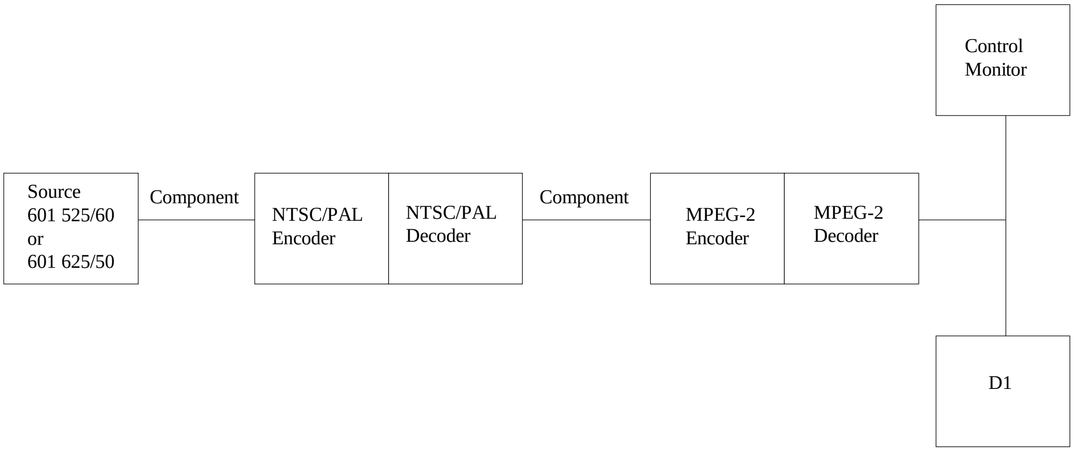


FIGURE 3 SEQUENCE PROCESSING

The sequences required for testing will be produced based on the block diagram shown in Figure 3. Rec. 601 Source component will be converted to Composite and back to Component (for HRC 4, 6 & 8 only) and passed through different MPEG-2 encoders at the various HRC's with the processed sequences recorded on a D1 VTR.

As a source video sequence passes through an HRC, it is possible that the resulting processed sequence has a number of scaling and alignment differences from the source sequence. To facilitate a common analysis of various objective quality measurement methods (referred to as models), Tektronix will normalize the processed sequences to remove the following deterministic differences that may have been introduced by a typical HRC:

- Global temporal frame shift (aligned to ±0 field error)
- Global horizontal/vertical spatial image shift (aligned to ±0.1 pixel)
- Global chroma/luma gain and offset ( normalization to no visible difference in alignment region)

Details of the normalization processing are given in Annex 3.

The processed and normalized sequences are then edited onto D1 test tapes using edit decision lists leading to the production of randomisation's distributed to each test facility for use in subjective testing sessions.

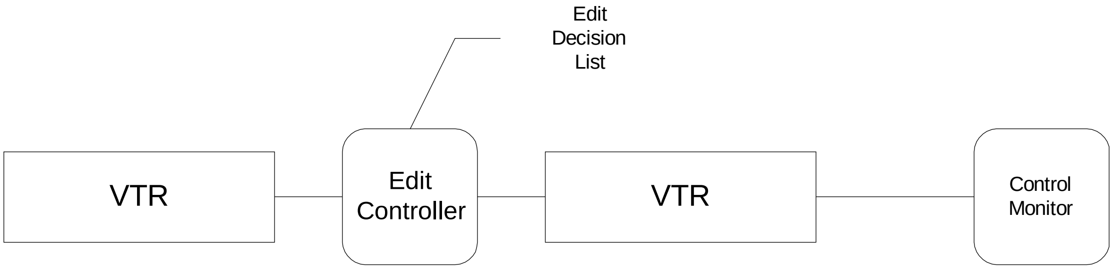


FIGURE 4 EDIT PROCESSING

### 3.3.3 RANDOMIZATIONS

The restrictions to the randomization rules in the determination of the order of trials in a test are listed here below:

- The 50% balance of distribution among Source to Processed (SP) vs. Processed to Source (PS) presentation order has to be balanced taking into account also the range of quality of HRC and the criticality of the video sequences. This requirement will be met in the following way:

  - The ILSC will rank order the HRCs from low quality to high quality for a given test. Let this rank ordering be given by (HRC 1, HRC 2, ...., HRC9).

  - The ILSC will rank order the source sequences from most critical (hardest to code) to least critical(easiest to code). Let this rank ordering be given by (source sequence 1, source sequence 2, ...., source sequence 10).

  - The following matrix will then be used to assign PS or SP ordering for the 90 clips in a given test.

| | HRC 1 | HRC 2 | HRC 3 | ... | HRC 9 |
|---|---|---|---|---|---|
| source sequence 1 | PS | SP | PS | ... | PS |
| source sequence 2 | SP | PS | SP | ... | SP |
| source sequence 3 | PS | SP | PS | ... | PS |
| . | . | . | . | ... | . |
| . | . | . | . | ... | . |
| . | . | . | . | ... | . |
| source sequence 10 | SP | PS | SP | | SP |

  The above assignment exactly balances PS and SP showings with respect to a given HRC and approximately balances the PS and SP showings with respect to a given source sequence (exact balance with respect to source sequence is not possible since there are only 9 HRCs in a test) and uniformly distributes the PS and SP orderings with respect to video quality.

- No two consecutive trials will present the same video sequence. (you can usually guarantee some minimum number of trials between presentations)
- Restrict number of consecutive trials based on identical Test Conditions (usually set to 1)
- Restrict maximum number of consecutive trial types PS and SP of the same type (usually set to 3),
- Try to ensure that no sequence is preceded by any other given sequence more than the minimum possible number of times (usually set to 1)

### 3.4 PRESENTATION STRUCTURE OF TEST MATERIAL

Due to fatigue issues, the sessions must be split into three sections: three about 30 minute viewing periods with two 20 minute (at least) breaks in between. This will allow for maximum exposure and best use of any one viewer.

Training trials (also called demonstration trials) will be recorded on separate tapes and run once per group of subjects at the very beginning of a test, assuming that all the test sessions forming the test are run at least in the same week.

Stabilisation trials (also called warm-up or reset trials) will be put before any test session without any noticeable interruption to the subjects.

The stabilisation phase will consist of 5 trials, selected from the actual material used for the test session, ensuring coverage of the full quality range.

Consequently, a typical session would consist of:

5 stabilization trials + 30 test trials
20 minute break
5 stabilization trials + 30 test trials
20 minute break
5 stabilization trials + 30 test trials

As an example this yields a group of up to 6 subjects evaluating 90 test trials at one time if two monitors are used. The subjects will remain in the same seating position for all 3 viewing periods.

As a compromise between the requirement to eliminate any contextual effect due to presentation order and the need to carry out the tests in a timely fashion, the following plan will be applied:

| Session Presentation Code | Session Presentation Order | | | Viewers | Labs |
|---|---|---|---|---|---|
| 1 | Session 1 | Session 2 | Session 3 | 1 – 6 | A, B |
| 2 | Session 2 | Session 3 | Session 1 | 7 – 12 | A, B |
| 3 | Session 3 | Session 1 | Session 2 | 13 – 18 (15) | A, B |
| 4 | Session 1 | Session 3 | Session 2 | 1 – 6 | C, D |
| 5 | Session 2 | Session 1 | Session 3 | 7 – 12 | C, D |
| 6 | Session 3 | Session 2 | Session 1 | 13 – 18 (15) | C, D |

## 4 VIEWING CONDITIONS

Viewing conditions should comply with those described Rec. ITU-R BT.500-8. An example of a viewing room is shown in Figure 5. Specific viewing conditions for subjective assessments in a laboratory environment are:

- Ratio of luminance of inactive screen to peak luminance: ≤ 0.02
- Ratio of the luminance of the screen, when displaying only black level in a completely dark room, to that corresponding to peak white*: ≅ 0.01
- Display brightness and contrast: set up via PLUGE (see Recommendations ITU-R BT.814 and ITU-R BT.815)
- Maximum observation angle relative to the normal**: $30^0$
- Ratio of luminance of background behind picture monitor to peak luminance of picture: ≅ 0.15
- Chromaticity of background: $D_{65}$ (0.3127, 0.3290)
- Peak screen luminance: 70 cd/$m^2$.
- Phosphor (x,y) chromaticities: R(0.640, 0.340), G(0.300, 0.600), B(0.150, 0.060) (these values are given in Rec. ITU-R BT.1361 and are close to both SMPTE-C and EBU values).

* It may become less than 0.01 when adjusted by PLUGE, but it is acceptable
**This number applies to CRT displays, whereas the appropriate numbers for other displays are under study.

The monitor size selected to be used in the subjective assessments is a 19" or 20" Professional Grade monitor. In the interest of uniformity of practice and because of the availability of 19" professional-grade monitors, the 19" condition supersedes the condition specified in Rec. ITU-R BT.1129-2 for 20" and over.

The viewing distance of 5H selected by VQEG falls in the range of 4 to 6 H, i.e. four to six times the height of the picture tube, compliant with Recommendation ITU-R BT.1129-2.

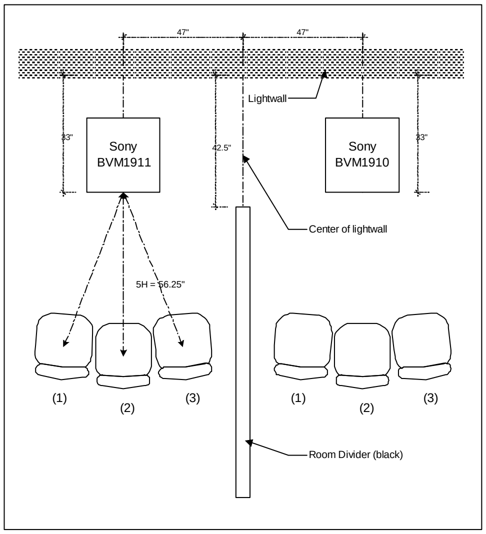


FIGURE 5 VIEWING ROOM AT THE CRC*

## 4.1 MONITOR DISPLAY VERIFICATION

Each subjective laboratory will undertake to ensure certain standards and will maintain records of their procedures & results, so that a flexible & usable standard of alignment 'objective' can be maintained.

It is important to assure the following conditions through monitor or viewing-environment adjustment:

- To make the display conditions uniform among different facilities, no aperture correction should be used.
- Monitor bandwidth should be adequate for the displayed format
- Focus should be adjusted for maximum visibility high-spatial-frequency information
- Purity (spatial uniformity of white field) should be optimized
- Geometry should be adjusted to minimize errors & provide desired overscan. The non-active video region is defined as:
  - ➢ the top 14 frame lines
  - ➢ the bottom 14 frame lines
  - ➢ the left 14 pixels
  - ➢ the right 14 pixels

Since the test sequences have the following size:
720 pixels by 486 frame lines for 525/60
720 pixels by 576 frame lines for 625/50

* As an example, this diagram shows the viewing room used for subjective tests at the Communications Research Centre (CRC).

and the active displayed region are
692 pixels by 458 frame lines for 525/60
692 pixels by 548 frame lines for 525/60,
the illustrated pattern below which has a border (WHITE) of 14 pixels or 14 lines width around the active video region (GRAY) will be recorded on the front portion of the test tapes.

```
 14                                        14
<----->< ------------------- 692 ------------------->< ---->

+-----------------------------------------------------------+   ---
|              [WHITE:235]                                  |    14
|      +-------------------------------------------+        |   ---
|      |                                           |        |
|      |                                           |        |
|      |                                           |        |
|      |           [GRAY:128]                      |        |
|      |         692 x 458 (525/60)                |        |  458/548
|      |         692 x 548 (625/50)                |        |
|      |                                           |        |
|      |                                           |        |
|      |                                           |        |
|      |                                           |        |
|      +-------------------------------------------+        |   ---
|                                                           |    14
+-----------------------------------------------------------+   ---
```

- Convergence should be optimized
- Black level set with PLUGE signal under actual ambient light conditions, as viewed from desired distance
- Luminance set to peak of 70 $cd/m^2$
- Greyscale tracking should be optimized for minimum variation between 10 and 100 IRE, with D6500 as target.
- Optical cleanliness should be checked.
- Video signal distribution system should be adequately characterized &adjusted.

In addition, it is useful to perform a test on the resolution of the screen (especially in high-luminance conditions) . [for further reading, see NIDL Display Measurement Methods available at <http://www.nta.org/SoftcopyQualityControl/MonitorReports/>: NIDL Monochrome Measurement Methods, Version 2.0 (1995); and NIDL Color Measurement Methods, Version 2.0 (1995).]:

==============
TEST PATTERNS:
==============

Three digital test patterns are available for use in monitor verification, which can be obtained by anonymous ftp from NIDL [instructions: (1) ftp to ftp.sarnoff.com; (2) input user name dvs; (3) input password 20dvs10]. These comprise six files (three tests in two optional formats). Each test is identified through its file name (pluge, tone, or vcal), and its format is identified through the extension (yuv or abk).

- Extension .yuv identifies the file as 720x480, 4:2:0 encoded, consecutive in Y, U, and V (all the Ys, then all the Us, then all the Vs).

- Extension .abk identifies the file as encoded according to the SMPTE 125M standard: that is, 720x486,4:2:2 encoded, and interleaved (Cb, Y0, Cr, Y1, etc.).

The three tests are as follows;

a) Pluge test (filename pluge), including white and the gray levels specified in Rec. ITU-R BT.814-1.

b) Gray scale test (filename tone), including nine squares with the gray levels 16, 48, 80, 112, 144, 176, 208, 235, and 255, all on a background of 170. Note that the value 255 may not be accessible in Rec 601 format, but that this point is removable from the data set.

c) Briggs test (filename vcal), including nine checkerboards at the cardinal screen positions (each pattern having a white-to-black-level difference of 6, and the patterns being at several different luminance levels). Only the center pattern need be incorporated in the quantitative test, with spot checks at the screen corners. The corrected versions of the original vcal.yuv which has legal Rec. 601 luminance values 16-235 only (vcalc.yuv 525/60 and vcal625c.yuv 625/50) is available at <http://www.commslab.gov.au/std/vqeg>. Using the Briggs test pattern, limiting resolution can be measured as follows:

- If the bottom two checkerboards are seen, the limiting resolution is supposed to be more than 180 samples per picture width (equal to or greater than 135 TVL).
- If the bottom three checkerboards are seen, the limiting resolution is supposed to be more than 360samples per picture width (equal to or greater than 270 TVL).
- If the bottom four checkerboards are seen, the limiting resolution is supposed to be more than 720 samples per picture width (equal to or greater than 540 TVL).

The test patterns will also be recorded on the front portion of the test tapes.

=======
REPORT:
=======

The following display conditions should be reported ([M]: mandatory, [O]: optional).

(1) Monitor specifications in the operational manual:
[M]Make and model
[M]CRT size (diagonal size of active area)
[M]Resolution (TVL)
[M]Dot-pitch (mm)
[M]Phospor chromaticities (x, y) for R, G, and B

(2) Display setup:
[M]Luminance of the inactive screen (in a normal viewing condition)
[M]Maximum obtainable peak luminance (in a dark room, measured after black-level adjustment before or during peak white adjustment)
[M]Luminance of the screen for white level (using PLUGE in a dark room)

What’s the reason for the distinction between the latter two measurements? Peak luminance can be varied by adjusting the contrast control and the maximum obtainable luminance level would be more than 100cd/m$^2$ in an ordinary case. The peak luminance should be adjusted to 70cd/m$^2$

as specified, and the latter item above corresponds to this adjusted luminance. If the maximum obtainable luminance level is almost equal to 70cd/m$^2$, some defects in displayed images would appear. The use of a monitor at its maximum obtainable luminance level or saturated level should be avoided. In order to ensure the appropriate monitor condition, it is required to report the maximum obtainable peak luminance.

[M]Luminance of the screen when displaying only black level (in a dark room)
[M]Luminance of the background behind a monitor (in a normal viewing condition)
[O/M]White balance and gamma (using gray scale test pattern in a dark room)

| video level | luminance(cd/m$^2$) | chromaticity (x, y) |
|---|---|---|
| 235 (white) | [O] | [O] |
| 208 | [O] | [O] |
| 176 | [O] | [M] |
| 144 | [O] | [O] |
| 112 | [O] | [O] |
| 80 | [O] | [O] |
| 48 | [O] | [M] |
| 16 (black) | [O] | [O] |

[O/M]Resolution (using Briggs test pattern in a dark room, report the perceived smallest check-sizes for each luminance level.)

| | left | center | right |
|---|---|---|---|
| top | [O] | [O] | [O] |
| center | [O] | [M] | [O] |
| bottom | [O] | [O] | [O] |

[M]Chromaticity of background (in a normal viewing condition)
[O]Phosphor chromaticities for R, G, and B (in a dark room)
[O]MTF (in a dark room)

3) Video distribution:
The objective of these measurements is simply that each lab be able to certify, in whatever way that they can, that their video distribution systems are essentially transparent.

[M] Block diagram of distribution system

If the video signal is supplied to monitors through analog interfaces, the following items also apply:

[ M] Frequency response of system
Although it would be preferable to use constant frequency signals such as an assortment of horizontal sine waves, it is certainly more practical to use the Multiburst signal that is internally generated by the D1.
The pattern displays bursts at 0.5, 1.0, 2.0, 4.0, 4.8 and 5.75 MHz.
The peak to peak amplitude at 1 MHz is to be normalized to 700mv on a waveform monitor, and the peak to peak amplitudes will then be measured at the other frequencies. The deviation at each frequency can then
be reported with respect to 1 MHz, and reported in dB where dB=20*log (measured / 700 ).
Where signal distribution is in RGB format, this measurement must be undertaken for each color channel.

[M] Interchannel gain difference
The internally generated 100% colorbars of the D1 can be used to ensure that the output levels of each channel are adjusted for unity for RGB, or to standard amplitudes for YPbPr. Report the peak to peak video level, in mv, for each color channel.

[O] Interchannel timing difference
This measurement may be a problem. An example how it could be made is given by the ATEL, using a HDTV bowtie pattern and subtraction on a waveform monitor to observe the timing

errors of the R and B channels relative to G. The worst case difference is reported. For example, if the R channel is 2ns early relative to G, and the B channel is 2ns late, then the timing error is 4ns. The interchannel delay is measured bypassing the D/A converter. Obviously, this will only work where RGB distribution is used. The problem is that no bowtie pattern in either 525 or 625 component interlaced is available.

[O] Nonlinearity
A 700mv ramp could be used for this measurement. The output of each channel (RGB) through the distribution system can be compared to the direct output of the green channel. Using subtraction on a waveform monitor, a system with zero nonlinearity will produce a perfectly horizontal trace. Report any deviations from the horizontal, in mv, for each color channel. In this case, the possible effect of the D/A conversion cannot be measured even if the test pattern is recorded on tape.

[O] Signal to noise ratio
The method used will depend upon the equipment available. Since the measurement is optional, the individual test labs should adopt whatever method is suitable, and report their method along with their results. An example is given by the ATEL, where a tangential method is applied using a 1780R waveform monitor to estimate the noise by observing any portion of a video signal with constant luminance. The white portion of a colorbar would be suitable. The sensitivity of this S/N measurement method is in the 55 to 60 dB range.

## 5 INSTRUCTIONS TO VIEWERS FOR QUALITY TESTS

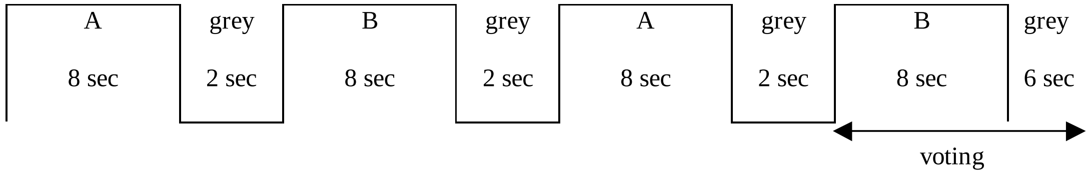
The following text (if necessary translated into the language of the respective country) shall be the instructions given to subjects. Slight modifications are allowed for labs where electronic devices are used instead of pen & paper.

"In this test, we ask you to evaluate the overall quality of the video material you see. We are interested in your opinion of the video quality of each scene. Please do not base your opinion on the content of the scene or the quality of the acting. Take into account the different aspects of the video quality and form your opinion based upon your total impression of the video quality.

Possible problems in quality include:

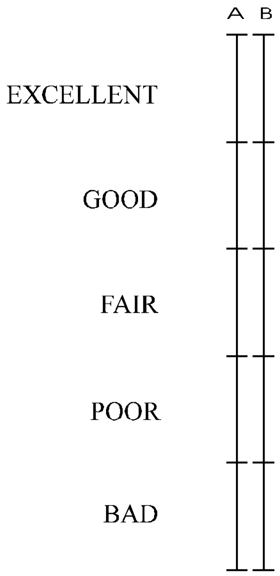
- poor, or inconsistent, reproduction of detail;
- poor reproduction of colours, brightness, or depth;
- poor reproduction of motion;
- imperfections, such as false patterns, or "snow".

The test consists of a series of judgement trials. During each trial, two versions of a single video sequence which may or may not differ in picture quality, will be shown in the following way:



"A" is the first version, "B" is the second version. Each trial will be announced verbally by number. The first presentation of a trial will be announced as "A", and the second as "B". This pair of presentations will then be repeated, thereby completing a single trial.

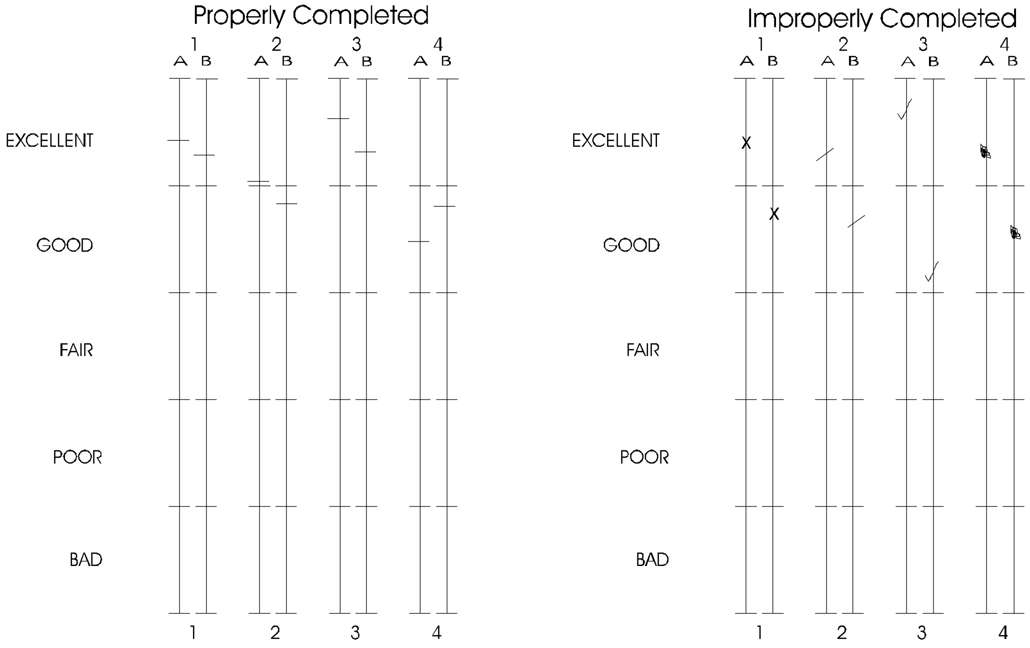
In judging the overall quality of the presentations, we ask you to use judgement scales like the samples shown below.



SAMPLE QUALITY SCALE

As you can see, there are two scales for each trial, one for the "A" presentation and one for the "B" presentation, since both the "A" and "B" presentations are to be judged.

The judgement scales are continuous vertical lines that are divided into five segments. As a guide, the adjectives "excellent", "good", "fair", "poor", and "bad" have been aligned with the five segments of the scales. You are asked to place a single horizontal line at the point on the scale that best corresponds to your judgement of the overall quality of the presentation (as shown in the example).



You may make your mark at any point on the scale which most precisely represents your judgement.

In making your judgements, we ask you to use the first pair of presentations in the trial to form an impression of the quality of each presentation, but to refrain from recording your judgements. You may then use the second pair of presentations to confirm your first impressions and to record your judgements in your Response Booklet.

We will now show you four demonstration trials. Please judge the quality using your training sheet**.**

DEMONSTRATION TRIALS PRESENTED AT THIS POINT

## 6 VIEWERS

Two different groups of 18 observers will be used in each laboratory, one for the high-quality test, the other one for the low-quality test. Only non-expert viewers will participate. The term non-expert is used in the sense that the viewers' work does not involve television picture quality and they are not experienced assessors. They must not have participated in a subjective quality test over a period of four months. All viewers will be screened prior to participation for the following:

- normal (20/20) visual acuity with or without corrective glasses (per Snellen test or equivalent)
- normal colour vision (per Ishihara test or equivalent)
- familiarity with the language sufficient to comprehend instruction and to provide valid responses using semantic judgement terms expressed in that language.

The results will be checked for completeness first. An observer is discarded if the number of failed votings exceeds one in one of the sessions. Additionally, the observers will be screened after the test as specified in sec. 2.3.1 of Annex 2 "Screening for DSIS, DSCQS and alternative methods except SSCQE method" of recommendation ITU-R BT.500-8. The viewers will be assigned to sub-groups which will see the test sessions in different orders (chapter 2.4). The screening will NOT be applied to these sub-groups but to the groups which participate in one test (e.g. 525/60, High Quality) as a whole.

Valuable results of at least 15 viewers are required. Consequently, an additional test is necessary if the number of viewers is reduced to less than 15 as a result of the screening.

## 7 DATA

### 7.1 RAW DATA FORMAT

Depending on the facility conducting the evaluations, data entries may vary, however the structure of the resulting data should be consistent among laboratories. An ASCII format data file should be produced with certain header information followed by relevant data pertaining to the ratings/judgements including the results of the stabilization trials see below:

In order to preserve the way in which data is captured, one file will be created with the following information:

RAW DATA

| Subject Number | SxHRCy | | SxHRCy | | SxHRCy | |
|---|---|---|---|---|---|---|
| | source | process | process | source | source | process |
| 111011 | 95.1 | 62.3 | 71.5 | 20.4 | 75.8 | 49.3… |
| 111021 | 88.6 | 60.4 | 75.1 | 21.2 | 77.0 | 51.3… |
| . | | | | | | |
| . | | | | | | |

The codes of the subject number have the following meaning:
1st digit: labs (see 3.3.1)

2nd digit: LQ(0) / HQ(1)
3rd digit: session order (see 3.4)
4th & 5th digit: subjects number (01-18)
6th digit: seat position (1-3, see FIGURE 5 in section 4)

All scene and HRC combination will be identified in the first row of the file. All these files should have extensions ".dat". This file will include the results of the stabilization trials. These also will be labeled. The files should be in ASCII format and/or Excel format.

## 7.2 SUBJECT DATA FORMAT

The purpose of this file is to contain all information pertaining to individual subjects who participate in the evaluating. The structure of the file would be the following:

| Subject<br>Number | Tape<br>Number | Month | Day | Year | Age | Gender* |
|---|---|---|---|---|---|---|
| 111011 | 01 | 02 | 12 | 98 | 25 | 2 |
| 111021 | 01 | 02 | 12 | 98 | 32 | 1 |

*Gender where 1=Male, 2=Female

## 7.3 DE-RANDOMIZED DATA

In a normal situation for the statistical analysis of data it is nice to have the data set sorted in order of scene and HRC combination. It is proposed that if possible each lab produce a data file with sorted data to resemble the following:

| | | | | SORTED SOURCE<br>DATA POINTS | | |
|---|---|---|---|---|---|---|
| Subject<br>Number | Tape | Age | Gender | S1HRC1 | S1HRC2 | S1HRC3.. |
| 111011 | 01 | 27 | 2 | 78.0 | 53.5 | 49.1 |

# 8 DATA ANALYSIS

The basic data analysis will include:

- MOS: Mean opinion score
- DMOS: Difference mean opinion score; Source - Processed
- Confidence interval

The lab to lab comparison will include some or all of the following measures:

- Pearson' Correlation Coefficient
- Spearman's Correlation Coefficient
- RMS error
- Weighted RMS Error
- Outlier Ratio of "outlier-points" to total points N
- Anova/Manova: Analysis of Variance - an inferential statistical technique used to compare differences between two or more groups with the purpose of making a decision that the independent variable influenced the dependent variable.

The high quality (HQ) and low quality (LQ) comparison should be carried out using ANOVA to examine contextual effect due to separating the test into HQ and LQ. Effects of "viewers group" and "context (HQ/LQ)" should be distinguished.

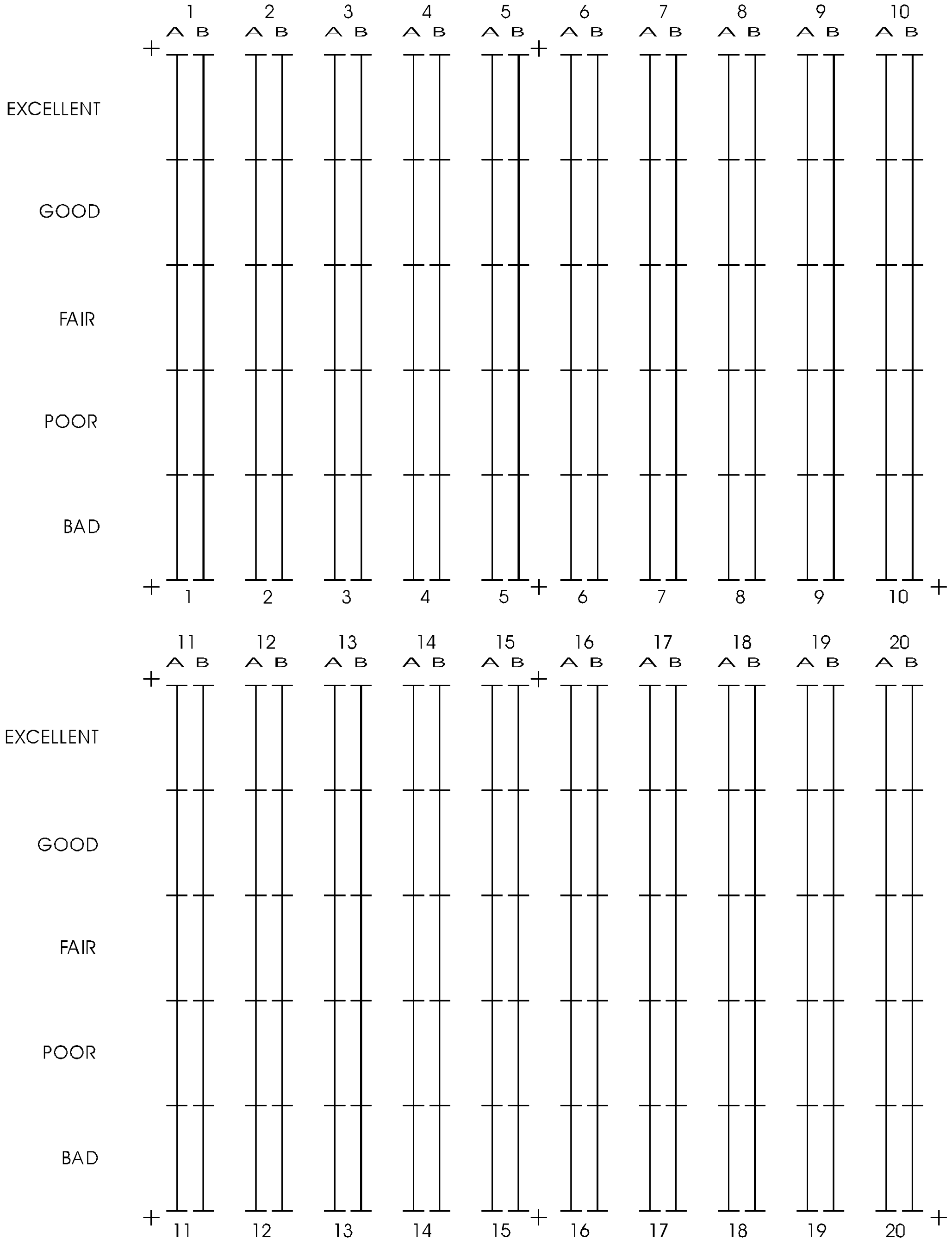
## 9 DEFINITIONS

- Source sequence is an unprocessed video segment recorded according to the sampling structure and colour format (4:2:2) of Rec. 601.
- Processed sequence: a source sequence encoded and decoded according a certain HRC.
- Test condition: any method of analog and/or digital processing of source sequences (also defined as Hypothetical Reference Circuits (HRCs).
- Training phase: include a detailed explanation of the test methodology and the test procedures and it usually includes also a generic dummy test session made with the same rule of the test sessions that the subjects are going to do, and made using the same test conditions but different source sequences (also defined as Demonstration phase).
- Stabilisation phase: is made of a certain number of test trials (typically five) duplicated from the test sessions, that cover the full range of quality; a stabilisation phase is inserted at the beginning of each test sessions; the votes collected during the stabilisation phase will not be included in the analysis.
- Test: is made up of one or more test sessions according to the length of time required to cover all the test conditions.
- Test session: is a group of test trials presented in a continuous period and organised not to exceed the maximum recommended viewing time of approximately 30 minutes.
- Trial: is presentation of a test condition.
- Test trial: is a trial for which subjects evaluate and the results are retained for analysis.
- Stabilisation trial: is a trial for which subjects evaluate and the results are Not retained for analysis.
- Training trial: is a trial for which subjects evaluate and the results are Not retained for analysis.
- Demonstration trials: is a trial which is presented solely to familiarise the subject with the structure of the trial.
- Test tapes are tapes containing randomized test trials.
- Edit decision lists are time code specifications for placement of test trials for the production of test tapes
- Contextual effects are fluctuations in the subjective rating of sequences resulting from the level of impairment in preceding sequences. For example, a sequence with medium impairment that follows a set of sequences with little or no impairment may be judged lower in quality than if it followed sequences with significant impairment.

## ANNEX 1: Sample page of response booklet

QT 1
DSCQS

| SUBJECT NO. | DATE / TIME | SESSION / SEAT | AGE / SEX | CITIZENSHIP | TAPE ORDER |
|---|---|---|---|---|---|
| | | | | | |



## ANNEX 2: Detailed information on HRCs

This part of the subjective test plan gives detailed information on the Hypothetical Reference Circuits (HRC) of TABLE 2.

**Choices of  625/50 processed sequences**

<table>
<tr><th>Seq. &amp;<br>HRC</th><th>Tree<br>(1)</th><th>Barcel.<br>(2)</th><th>Harp<br>(3)</th><th>Mov.<br>(4)</th><th>Canoa<br>(5)</th><th>F1 car<br>(6)</th><th>Fries<br>(7)</th><th>Hor. S.<br>(8)</th><th>Rugby<br>(9)</th><th>Table<br>(10)</th><th>Flower<br>(11)</th><th>M&amp;C<br>(12)</th></tr>
<tr><td>Ref</td><td colspan="12">Reference</td></tr>
<tr><td>1</td><td colspan="12">1.1</td></tr>
<tr><td>2</td><td colspan="12">2.1</td></tr>
<tr><td>3</td><td colspan="12">3.1</td></tr>
<tr><td>4</td><td colspan="12">4.1</td></tr>
<tr><td>5</td><td colspan="6">5.2</td><td colspan="6">5.1</td></tr>
<tr><td>6</td><td colspan="6">6.6</td><td colspan="6">6.4</td></tr>
<tr><td>7</td><td colspan="12">7.2</td></tr>
<tr><td>8</td><td colspan="4">8.2</td><td colspan="4">8.3</td><td colspan="4">8.5</td></tr>
<tr><td>9</td><td colspan="6">9.3</td><td colspan="6">9.4</td></tr>
<tr><td>10</td><td colspan="12">10.2</td></tr>
<tr><td>13</td><td colspan="2">13.2</td><td colspan="4">13.1</td><td colspan="6">13.2</td></tr>
<tr><td>14</td><td colspan="6">14.2</td><td colspan="6">14.3</td></tr>
</table>

**Production of 625/50 processed sequences**

| Version | HRC | Coding parameters | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Bit rate [Mbit/s] | Method | Encoder | Resol H pixel | GOP | Mode (1) | Note (2) |
| 14.1 | 14 | 2 | MP@ML | Coder 1 | 528 | N=12 M=3 | 1 | - |
| 14.2 | 14 | 2 | MP@ML | Coder 2 | 3/4 | N=12 M=3 | Auto | - |
| 14.3 | 14 | 2 | MP@ML | Coder 3 | 544 | N=12 M=3 | 3 | |
| | | | | | | | | |
| 13.1 | 13 | 2 | SP@ML (3) | Coder 1 | 528 | N=12 M=1 | 1 | - |
| 13.2 | 13 | 2 | SP@ML (3) | Coder 2 | 3/4 | N=12 M=1 | Auto | - |
| | | | | | | | | |
| 10.1 | 10 | 4.5 | MP@ML | Coder 1 | 704 | N=12 M=3 | 1 | - |
| 10.2 | 10 | 4.5 | MP@ML | Coder 1 | 704 | N=12 M=3 | 2 | - |
| 10.3 | 10 | 4.5 | MP@ML | Coder 2 | full | N=12 M=3 | Auto | - |
| 10.4 | 10 | 4.5 | MP@ML | Coder 3 | 720 | N=12 M=3 | 3 | - |
| | | | | | | | | |
| 9.1 | 9 | 3 | MP@ML | Coder 1 | 704 | N=12 M=3 | 1 | - |
| 9.2 | 9 | 3 | MP@ML | Coder 1 | 704 | N=12 M=3 | 2 | - |
| 9.3 | 9 | 3 | MP@ML | Coder 2 | full | N=12 M=3 | Auto | - |
| 9.4 | 9 | 3 | MP@ML | Coder 3 | 720 | N=12 M=3 | 3 | - |
| | | | | | | | | |
| 8.1 | 8 | 4.5 | MP@ML | Coder 1 | 704 | N=12 M=3 | 1 | PAL 1 |
| 8.2 | 8 | 4.5 | MP@ML | Coder 1 | 704 | N=12 M=3 | 1 | PAL 2 |
| 8.3 | 8 | 4.5 | MP@ML | Coder 2 | full | N=12 M=3 | Auto | PAL 4 |
| 8.4 | 8 | 4.5 | MP@ML | Coder 3 | 720 | N=12 M=3 | 3 | PAL 1 |
| 8.5 | 8 | 4.5 | MP@ML | Coder 3 | 720 | N=12 M=3 | 3 | PAL 2 |
| | | | | | | | | |
| 7.1 | 7 | 6 | MP@ML | Coder 1 | 704 | N=12 M=3 | 1 | - |
| 7.2 | 7 | 6 | MP@ML | Coder 2 | full | N=12 M=3 | Auto | - |
| 7.3 | 7 | 6 | MP@ML | Coder 3 | 720 | N=12 M=3 | 3 | - |
| | | | | | | | | |
| 6.1 | 6 | 8 | MP@ML | Coder 1 | 704 | N=12 M=2 | 1 | PAL 1 |
| 6.2 | 6 | 8 | MP@ML | Coder 1 | 704 | N=12 M=2 | 1 | PAL 2 |
| 6.3 | 6 | 8 | MP@ML | Coder 2 | full | N=12 M=2 | Auto | PAL 4 |
| 6.4 | 6 | 8 | MP@ML | Coder 2 | full | N=12 M=2 | Film | PAL 4 |
| 6.5 | 6 | 8 | MP@ML | Coder 3 | 720 | N=12 M=3 | 3 | PAL 1 |
| 6.6 | 6 | 8 | MP@ML | Coder 3 | 720 | N=12 M=3 | 3 | PAL 2 |
| | | | | | | | | |
| 5.1 | 5 | 8 | MP@ML | Coder 1 | 704 | N=12 M=2 | 1 | - |
| | | 4.5 | MP@ML | Coder 3 | 720 | N=12 M=3 | 3 | - |
| 5.2 | 5 | 8 | MP@ML | Coder 2 | Full | N=12 M=2 | Auto | - |
| | | 4.5 | MP@ML | Coder 2 | Full | N=12 M=3 | Auto | - |
| | | | | | | | | |
| 4.1 | 4 | 19 | 4:2:2P@ML | Coder 3 | 720 | N=12 Low Delay | 3 | - |
| | | | PAL | | | | | PAL 3 |
| | | 19 | 4:2:2P@ML | Coder 3 | 720 | N=12 Low Delay | 3 | |
| | | | PAL | | | | | PAL 3 |
| | | 12 | MP@ML | Coder 3 | 720 | N=12 M=3 | 3 | |
| | | | | | | | | |
| 3.1 | 3 | 50<br>7th gen. | 4:2:2P@ML | Coder 3 | 720 | Intra | - | Shift (4) |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |
| 2.1 | 2 | 19 | 4:2:2P@ML | Coder 3 | 720 | N=12<br>Low Delay | 3 | Shift<br>1 line down<br>2Y pixel right |
| | | 19 | 4:2:2P@ML | Coder 3 | 720 | N=12<br>Low Delay | 3 | Shift<br>1 line up<br>2Y pixel left |
| | | 12 | MP@ML | Coder 3 | 720 | N=12 M=3 | 3 | |
| | | | | | | | | |
| 1.1 | 1 | 4th gen.<br>(5) | BetacamSP | Coder 3 | n/a | n/a | - | |

Legenda:

(1) 1: Video mode forced 2: Film mode enabled
3: interlaced 4: non interlaced

(2) PAL 1 → PAL Coder **D-5** & PAL Decoder **Vistek**
PAL 2 → PAL Coder **Vistek** & PAL Decoder **D-5**
PAL 3 → PAL Coder **Vistek** & PAL Decoder **Sony**
PAL 4 → PAL Coder **D5** & PAL Decoder **Digital Betacam**

(3) MP@ML encoder used without B frames.

(4) Shift performed in multigeneration

original--> codec-->( 1° generation) D1-->shift -->codec-->( 2° generation) D1-->shift and so on..

ORI --> no shift (useless) -->1° generation
1° generation --> 1 line down --> 2° generation
2° generation --> 1 line down --> 3° generation
3° generation --> 2 luma pixel right --> 4° generation
4° generation --> 1 line up --> 5° generation
5° generation --> 2 luma pixel left --> 6° generation
6° generation --> 1 line up --> 7° generation

in this way no shift vertical or horizontal between ORI and 7th generation occurs.

(5) BetacamSP multigeneration has been performed using a Betacam SP brand new tape recorded on a Sony BVW-75P VTR and D-1 as intertape.

**Choices of  525/60 processed sequences**

<table>
<tr><th>Seq. &amp;<br>HRC</th><th>Ballon<br>(13)</th><th>New Y.<br>(14)</th><th>M&amp;C<br>(15)</th><th>Betes<br>(16)</th><th>Le P.<br>(17)</th><th>Aut.<br>(18)</th><th>Foot<br>(19)</th><th>Sail<br>(20)</th><th>Susie<br>(21)</th><th>Temp.<br>(22)</th><th>Table<br>(23)</th><th>Flower<br>(24)</th></tr>
<tr><td>2</td><td colspan="12">2.1</td></tr>
<tr><td>3</td><td colspan="12">3.1</td></tr>
<tr><td>4</td><td colspan="12">4.1</td></tr>
<tr><td>5</td><td colspan="6">5.1</td><td colspan="6">5.2</td></tr>
<tr><td>6</td><td colspan="12">6.2</td></tr>
<tr><td>7</td><td colspan="12">7.1</td></tr>
<tr><td>8</td><td colspan="12">8.2</td></tr>
<tr><td>9</td><td colspan="6">9.1</td><td colspan="6">9.2</td></tr>
<tr><td>10</td><td colspan="12">10.3</td></tr>
<tr><td>13</td><td colspan="6">13.3</td><td colspan="6">13.1</td></tr>
<tr><td>14</td><td colspan="4">14.1</td><td colspan="4">14.4</td><td colspan="4">14.3</td></tr>
</table>

**Production of 525/60 processed sequences**

| Ver sion | HR C | Coding parameter | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Bit rate [Mbit/s] | Method | Encoder | Resol H pixel | GOP | Mode (1) | Note (2) |
| 14.1 | 14 | 2 | MP@ML | Coder 1 | 528 | N=15 M=3 | 1 | - |
| 14.2 | 14 | 2 | MP@ML | Coder 3 | 544 | N=15 M=3 | 3 | |
| 14.3 | 14 | 2 | MP@ML | Coder 2 | 3/4 | N=15 M=3 | Video | |
| 14.4 | 14 | 2 | MP@ML | Coder 2 | 3/4 | N=15 M=3 | Film | |
| | | | | | | | | |
| 13.1 | 13 | 2 | SP@ML (3) | Coder 1 | 528 | N=15 M=1 | 1 | - |
| 13.2 | 13 | 2 | SP@ML (3) | Coder 2 | 3/4 | N=15 M=1 | Video | - |
| 13.3 | 13 | 2 | SP@ML (3) | Coder 2 | 3/4 | N=15 M=1 | Film | - |
| | | | | | | | | |
| 10.1 | 10 | 4.5 | MP@ML | Coder 1 | 704 | N=15 M=3 | 1 | - |
| 10.2 | 10 | 4.5 | MP@ML | Coder 3 | 720 | N=15 M=3 | 3 | - |
| 10.3 | 10 | 4.5 | MP@ML | Coder 2 | full | N=15 M=3 | Video | - |
| | | | | | | | | |
| 9.1 | 9 | 3 | MP@ML | Coder 1 | 704 | N=15 M=3 | 1 | - |
| 9.2 | 9 | 3 | MP@ML | Coder 3 | 720 | N=15 M=3 | 3 | - |
| 9.3 | 9 | 3 | MP@ML | Coder 3 | 720 | N=15 M=3 | 4 | - |
| 9.4 | 9 | 3 | MP@ML | Coder 2 | full | N=15 M=3 | Video | - |
| | | | | | | | | |
| 8.1 | 8 | 4.5 | MP@ML | Coder 1 | 704 | N=15 M=3 | 1 | NTSC 1 |
| 8.2 | 8 | 4.5 | MP@ML | Coder 3 | 720 | N=15 M=3 | 3 | NTSC 1 |
| | | | | | | | | |
| 7.1 | 7 | 6 | MP@ML | Coder 1 | 704 | N=15 M=3 | 1 | - |
| 7.2 | 7 | 6 | MP@ML | Coder 3 | 720 | N=15 M=3 | 3 | - |
| 7.3 | 7 | 6 | MP@ML | Coder 2 | full | N=15 M=3 | Video | - |
| | | | | | | | | |
| 6.1 | 6 | 8 | MP@ML | Coder 1 | 704 | N=15 M=3 | 1 | NTSC 1 |
| 6.2 | 6 | 8 | MP@ML | Coder 3 | 720 | N=15 M=3 | 3 | NTSC 1 |
| | | | | | | | | |
| 5.1 | 5 | 8 | MP@ML | Coder 1 | 704 | N=15 M=3 | 1 | - |
| | | 4.5 | MP@ML | Coder 3 | 720 | N=15 M=3 | 3 | - |
| 5.2 | 5 | 8 | MP@ML | Coder 2 | Full | N=15 M=2 | Video | - |
| | | 4.5 | MP@ML | Coder 2 | Full | N=15 M=3 | Video | - |
| | | | | | | | | |
| 4.1 | 4 | 19 | 422P@ML | Coder 3 | 720 | (5) | | |
| | | | NTSC | | | | | NTSC 2 |
| | | 19 | 422P@ML | Coder 3 | 720 | (5) | | |
| | | | NTSC | | | | | NTSC 2 |
| | | 12 | MP@ML | Coder 3 | 720 | N=15 M=3 | | |
| | | | | | | | | |
| 3.1 | 3 | 50 | 422P@ML | Coder 3 | 720 | Intra | - | Shift (4) |
| | | | | | | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2.1 | 2 | 19 | 422P@ML | Coder 3 | 720 | (5) | | Shift<br>1 line down<br>2Y pixel right |
| | | 19 | 422P@ML | Coder 3 | 720 | (5) | | Shift<br>1 line up<br>2Y pixel left |
| | | 12 | MP@ML | Coder 3 | 720 | N=15 M=3 | | |

Legenda:

(1) 1:Video mode forced
2: Film mode enabled
3: Interlaced
4: Non interlaced

(2) NTSC 1 → NTSC Coder **Vistek** & NTSC Decoder **Vistek**
NTSC 2 → NTSC Coder **Vistek** & NTSC Decoder **Sony**

(3) MP@ML encoder used without B frames.

(4) Shift performed in multigeneration

original--> codec-->( 1° generation) D1-->shift -->codec-->( 2° generation) D1-->shift and so on..

ORI --> no shift (useless) -->1° generation
1° generation --> 1 line down --> 2° generation
2° generation --> 1 line down --> 3° generation
3° generation --> 2 luma pixel right --> 4° generation
4° generation --> 1 line up --> 5° generation
5° generation --> 2 luma pixel left --> 6° generation
6° generation --> 1 line up --> 7° generation

in this way no shift vertical or horizontal between ORI and 7th generation occurs.

(5) Low delay mode I plus adaptive slice of P

## ANNEX 3: Normalization Processing

### Normalization of Video Test Sequences and Results Summary

18 June, 1999
Jerry Lu, Norm Franzen, Wilfried Osberger
Tektronix, Inc.

This document is revised from the stand-alone report, "Normalization of Video Test Sequences for The Video Quality Experts Group." To fit the Subjective Test Plan, the results section of this document only contains a condensed result summary. A complete list of the normalization parameter results is now only available as Section 4.3 in the stand-alone version. Except for this difference, the two versions are essentially identical.

## 1. Introduction

This document first describes the objectives and procedures for video sequence normalization as required by the Video Quality Expert Group (VQEG) Objective and Subjective Test Plans. At the end of the document, it summarizes the results obtained during the normalization process.

The VQEG test video data were processed by various digital and analog HRC's. In producing impairments to the test sequences, many of the HRC's also created spatial and temporal shifts as well as gain and level changes. To facilitate the evaluation of objective quality models on a common basis, the normalization procedure removes the spatial/temporal shifts and gain/level changes from the test sequences relative to the source sequences.

The test data to be normalized were selected by the VQEG Independent Labs and Selection Committee (ILSC). The test data set includes 20 sequences processed by 16 HRC's which covers material that has been compressed at bit-rates ranging from 768 Kbit/s to 36 Mbit/s. It also includes video sequences that have passed through multiple impairment sources to simulate video signal changes over a distribution chain. The average PSNR of the sequences ranges from 18.0dB to 47.0dB.

The normalization procedure consisted of the following main steps:

1) preliminary sequence screening,
2) temporal alignment,
3) spatial shift estimation and correction,
4) spatial alignment verification,
5) gain and level estimation and correction,
6) verification of constant spatial shift, and
7) final visual verification.

The normalization procedure was applied to all the VQEG test sequences sent by the Canadian Research Centre to Tektronix from September 1998 through January 1999. The actual normalization correction values and the verified accuracy are summarized in the results section (Section 4). Also included in the normalization results are the analysis and observations from testing video sequences against other VQEG requirements, such as picture frame cropping. Those sequences that were found to be beyond the specified variation limits are also reported.

The methods and results are summarized in several sections. Section 2 defines the video normalization objectives. Section 3 describes the normalization procedure and specific methods, as well as the measures to ensure the robustness of the spatial and temporal alignments. Both automatic and manual operations are described in this section. Section 4 gives a summary of the normalization results. The complete set of the normalization parameters is provided in the complete document, “Normalization of Video Test Sequences for the Video Quality Experts Group.”

## 2. Video Normalization Objectives

Video normalization aims to accomplish the following two specific goals as defined by VQEG’s Objective and Subjective Test Plans:

1) registration of the test video sequences with the source sequences, both spatially and temporally, and
2) removal of two linear distortions that may exist in HRC impaired sequences: gain and level errors.

As implied by the VQEG requirements, it is assumed that there are only temporal delays, horizontal and vertical shifts, and linear distorion in the impaired sequences. It has been assumed that there is no scaling, rotation, and time-varying spatial shift. To define a consistent scope of the task, we specify the video normalization in more detail as follows:

1) chroma and luma spatial alignment will be applied to the Y, Cb, Cr channels independently,
2) chroma/luma gain and level will be corrected,
3) cropping and spatial misalignments will be assumed to be global, i.e. constant throughout the sequence, and
4) frame drop will be detected.

VQEG also defined a set of normalization accuracy requirements as in both the Objective Test Plan and the Subjective Test Plan. The normalization should completely correct any temporal shift. It should correct a spatial shift to within ±0.1 pixels, and adjust gain and level to no visible difference with respect to the source video.

The VQEG video sequences are in 4:2:2 format, and the sequences have been edited to contain the following syntax, which includes the alignment pattern to facilitate the spatial and temporal alignment and frame drop detection:

$AL(1sec) + VIDEO\ (10fr{+}8sec{+}10fr) + AL(1sec)$,

AL: alignment pattern.

## 3. Normalization Algorithm and Procedure

### 3.1 Preliminary sequence screening

There are two preliminary screening procedures. Both procedures are used to detect unintended errors that do not belong to the intended HRC impairments. These include D-1 medium errors, recording errors, playback errors, and local file transfer errors, etc. The first procedure involves a visual inspection of the un-normalized test video on the original D-1 tapes, which was done whenever possible. In the second procedure, we review the output from a video scene characterization algorithm that generates a unique signature curve for each sequence. A random error can be identified from the scene signature of the sequence. This is done after each video sequence, originally recorded in D-1 video tapes, is transferred over to computer disk files. Both the source and HRC-impaired video are tested using this screening process.

**3.2 Temporal alignment and frame drop detection**

A video sequence, after being transferred from a D-1 source tape through the Ethernet LAN to the workstation hard disk, can contain 1-4 more frames before and after the sequence because of the inaccuracy in D-1 playback synchronization with the frame buffer. As a result, there can be a temporal shift between the source and a test sequence.

The temporal shift detection is performed separately from the spatial shift detection. It is done by a scene break detector that locates the scene break between the two alignment pattern segments and the main segment of the test sequences. The scene break detection yields two quantities for each sequence, $I_{b1}$ and $I_{b2}$, which are the frame indexes to the two scene break locations. The temporal shift between a test and source is given by:

$$\textit{temporal shift} = I_{b1,test} - I_{b1,src}$$

The index values for $I_{b1,src}$ are known for the source video according to the test sequence syntax: 25 for 625/50, and 30 for 525/60 formatted sequences. A frame drop or frame count error occurs if the frame count between the scene breaks gives a value other than the expected length: 220 frames for 625/50, and 260 frames for 525/60 sequences.

$$\textit{length of the video content segment} = I_{b2,test} - I_{b1,test}$$

The same operation is performed on all the source video as well for verification.

**3.3 Spatial shift estimation and correction**

The spatial shift of a test sequence with respect to the corresponding source sequence is estimated over the first segment containing the alignment pattern. The correction algorithm uses the misalignment estimated in the horizontal and vertical directions to re-align the sequence. The same correction is applied to every frame in the sequence.

To control the alignment accuracy, an independent alignment verification is inserted between the shift estimation and correction. The verification, which determines when an estimate is good enough, is described in section 3.4. The relations of these processing functions are illustrated in Figure 1.

The alignment pattern contains alternating rectangular blocks in each of the individual channels: Y, Cb, and Cr. The spatial shift estimator uses the pattern to detect and measure the misalignment in each channel. The details of the alignment pattern can be found in the Appendix.

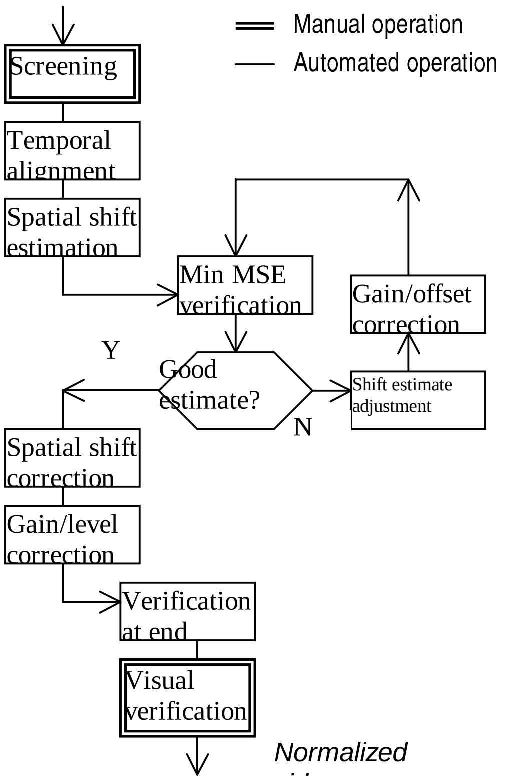


Figure 1. Normalization procedures

**3.3.1** ***Spatial shift estimation***

Spatial shift estimation uses a coarse-to-fine image registration method that can be described in two steps. In the first step, shifts in both directions are estimated to single pixel accuracy, providing an initial estimate of the spatial shift at a coarse level. Sub-pixel shift estimation is calculated in the second step to refine the previous shift detection results. Phase correlation [1] is used for the spatial shift estimation in both X and Y directions. In the second step, the phase correlation function is used in conjunction with a group delay estimation to refine the coarse shift estimate to sub-pixel accuracy [3]. The estimates are averaged over as many as ten frames to reduce the noise level and possible estimation bias in the presence of signal-dependent noise. This type of noise is common in digital video that has been compressed by MPEG encoders and many other compression techniques.

**3.3.2** ***Spatial shift correction***

Spatial shifts in a test sequence are corrected by moving each of the frames in the opposite direction of the misalignment by the same amount. Because of the possibility that a test sequence may have been shifted differently in the three channels, the correction must be applied to each of them separately. The corrected Y, Cr, and Cb are then reassembled into an aligned 4:2:2 test video.

The correction procedure applies a shift operation depending on the type of the misalignment. If an integer valued horizontal shift or an even-integer valued vertical shift is found, the correction is done by adding constants to the coordinates for each of the pixels in the picture frames. In the case of a subpixel shift, including an odd integer vertical shift, an interpolation is made by resampling a *sinc* interpolation function multiplied by a raised cosine window function.

When a shift is being corrected, the pixels shifted into the picture frame are assigned a 4:2:2 black pixel value of (Y, Cr, Cb) = ( 16, 128, 128 ), resulting in a wide black border at one or two edges of the frame. For instance, if the original test video contains a shift of eight pixels toward left with respect a reference, the corrected video will have a black border eight pixels wide on the left edge. It must be noted that there is a cropping effect caused by any shift correction operation. In effect, in situations where the amounts of the corrections differ in the three luminance and chrominance channels, the effective cropping is always the maximum of the corrections that were to be applied. This suggests that the amount of spatial shifts existing in the test sequences must also be checked against the limit VQEG imposed on the cropping. These results are given in Table 1 of the results section (Section 4).

### 3.4 Spatial alignment verification

Verification and refinement of the spatial alignment is performed using an independent procedure to check for optimal spatial alignment after normalization. Since the true value of the shift is unknown, a criterion is needed for checking for optimal alignment. A common criterion for finding an optimal alignment is the minimum mean square error (MSE), that is, the optimal correction is the shift operation that yields the minimum MSE.

This verification process is inserted between the shift estimation and the shift correction. It is responsible for two tasks. One is to determine if an additional adjustment is needed from the current shift estimate that would yield lower MSE; the next is to make the corresponding modification to the current shift estimate to obtain a better MSE output. These two steps are iterated until the minimum MSE is found. A sequential search is used in each of the two X-Y directions with a step size of 0.1 pixels to prove the optimality of the current estimate, and if not, to seek a fine adjustment to the shift estimate. Although the phase correlation shift estimator is accurate in general, the verification process is allowed to search as long as there is a better estimate or a maximum number of iterations is exceeded. The verification process controls the accuracy and improves the shift estimation in the event of severe distortions in the alignment pattern area.

### 3.5 Gain and level estimation and correction

Once a test sequence is aligned temporally and spatially with its source sequence, the linear gain and level distortions can be estimated. The estimator compares the test video with its source and fits a linear model to the correlation between the source video and the test video. The gain and level are the coefficients of the model. The estimation is performed using the first alignment patterns of the test video sequence and its source video sequence. The best fit of the model to the data yields the estimated values of the gain and level for the sequence. This operation is also performed on each of the Y, Cb, and Cr channels independently, producing gain and level estimates for each channel. Using minimum MSE as the best fit criterion, the gain and level are computed as follows:

$G = ( R_{xy} - R_x R_y )/( R_{xx} - R_x R_x )$, and
$L = R_y - G\, R_x$,

where $R_{xy}$ is the cross correlation function:

$R_{xy} = (1/NM)\ \Sigma\Sigma\ X(i, j)\ Y(i, j)$, and
$R_x = (1/NM)\ \Sigma\ \Sigma X(i, j),\ \ R_y = (1/NM)\ \Sigma\Sigma\ Y(i, j)$.

$N$ and $M$ are the dimensions of a video frame. $G$ and $L$ are the gain and level estimates, respectively.

The gain and level correction takes place when the alignment is complete, as shown in the diagram in Figure 1. Correction of the gain and level of a video frame uses the linear operation:

$I_{corrected} = (I - L)/G$ .

As the gain and level are estimated separately for each luminance and chrominance channel, the correction is applied independently.

### 3.6 Verification of constant spatial shift

The verification procedure defined above is applied again to test for any residual shift in the end alignment pattern when the spatial correction for an entire sequence is complete. A residual suggests a difference of the spatial shifts between the front and end of the test sequence. This is possible if different segments of the sequence are processed separately by a HRC, in which case the frame drop test described earlier in section 3.2 becomes invalid. It is also possible that a time-varying shift may have been caused by a HRC. In either case, the temporal and spatial alignment as well as the gain and level correction may no longer give valid results. These sequences are marked in Table 1 in the results section.

### 3.7 Final visual verification

Visual verification is the final step to verify the results of the normalization. In addition, it probes any exceptional cases where the automated procedure may not function as intended. The main items that were visually checked were the temporal and spatial shift residuals, and inconsistent spatial shifts.

The following visual examinations were performed:

1) Visual check on the difference between the beginning and ending alignment patterns to verify the status of incoherent spatial shift.
2) Viewing of a normalized sequence, its source, and the un-normalized sequence near the locations of the two scene breaks at the start and end of the test video content. This is to verify the spatial and temporal alignment as well as the gain and level correction.
3) Viewing of the normalized sequences on video monitors by multiple television experts.

## 4. Normalization Results

### 4.1 Overview

The 320 VQEG test sequences that have been normalized and verified all meet the VQEG requirements of 0.1-pixel maximum alignment error and zero field temporal error. Confirmed by the visual verification of the normalized sequences, all meet the criterion of no-visible difference in gain and level, as specified in the Subjective Test Plan, with a few exceptions that are noted in section 4.2.

The timeline of the normalization tasks are summarized in the table below.

| Normalization start (data received) | Normalized data completed and delivered |
|---|---|
| 09/09/98 (HRC's 1-10,13,14) | 10/25/98-11/22/98 (in D-1, Exabyte, DLT) ** |
| 12/04/98 (HRC's 15,16) | 12/17/98 (in D-1, Exabyte, DLT) * |
| 01/06/99 (HRC's 11,12) | 02/08/99 (in Exabyte, DLT) * |

* Exabyte copies to NTIA, Sarnoff, CRC, CPqD, and IfN. DLT copies to KPN, NHK, and EPFL. D-1 to CRC.

** NHK contributed duplication work for the 525/60 D-1 set.

### 4.2 Summary of several conditions related to the video normalization

***Table 1. Sequences with conditions near or beyond the limits specified by the VQEG Objective Test Plan or the Subjective Test Plan***

| **Sequence, HRC, Format** | **Conditions** |
|---|---|
| Src19,20,21,22, HRC5, 525/60 | -16 pixels of vertical shift, the net effect of cropping of at least 16 pixels at the top of the frames. |
| Src13, HRC15, 525/60 | -16 pixels of horizontal shift, the net effect of cropping of at least 16 pixels at right edge of the frames after shift correction. |
| Src13, HRC16, 525/60 | -16 pixels of horizontal shift, the net effect of cropping of at least 16 pixels at the right edge of the frames after shift correction. |
| Src19, Src21, HRC4, 525/60 | Slightly red tone in the corrected sequences. |
| Src14, 16, 18, 20, 22, HRC6, 525/60 | Slightly red tone in the corrected sequences. |
| Src21, HRC7, Src19, HRC8 | Slightly red tone in the corrected sequences. |
| Src1-10, HRC1, 625/50 | -13.3 pixels of horizontal shift, the net effect of cropping of at least 13.3 pixels at the right edge of the frames after shift correction. |
| Src3,4,5,6, HRC5, 625/50 | Unequal amount of spatial shift on the two alignment template segments. |
| Src3,4,5,6,8,9 HRC7, 625/50 | Unequal amount of spatial shift on the two alignment template segments. |
| Src7,8, HRC8, 625/50 | Unequal amount of spatial shift on the two alignment template segments. |
| Src3,4,5,6, HRC9, 625/50 | Unequal amount of spatial shift on the two alignment template segments. |

| | |
|---|---|
| Src8,9,10, HRC13, 625/50 | Unequal amount of spatial shift on the two alignment template segments. |
| Src3,4,5,6, HRC14, 625/50 | Unequal amount of spatial shift on the two alignment template segments. |
| Src1, HRC15, 625/50 | Unequal amount of spatial shift on the beginning alignment segment and the main video segment. Alignment was done based on visual observation to align the video segment only. |
| Src1, HRC16, 625/50 | Unequal amount of spatial shift on the beginning alignment segment and the main video segment. Alignment was done based on visual observation to align the video segment only. |

Note: In all the sequences with unequal amount of spatial shift on the two alignment stripe segments, the difference in shifts is 0.4 pixels.

**Table 2. Other conditions in the original data**

| **Sequence, HRC, Format** | **Conditions** |
|---|---|
| Src13, HRC1 525/60 | Line dropouts in frames 261 and 263. First field blank |
| Src19, HRC1 525/60 | There are one or two frames with blurred stripe |
| Src20, HRC1 525/60 | Line dropouts in frame 77 |
| Src20, HRC4 525/60 | Digital dropouts |
| Src16, HRC6 525/60 | Line dropout in frame 31 |
| Src7, all HRC’s 625/50 | Top right corner cropped |
| Src10, HRC4, 625/50 | Digital dropouts |

## 5. References

[1] A. Murat Tekalp, Digital Video Processing, Prentice-Hall, 1995.
[2] A. V. Oppenheim, R. W. Schafer, Discrete-Time Signal Processing, Prentice-Hall, 1989.
[3] Phase correlation [1] is used as the estimation algorithm in the first step, which is an algorithm only for estimating at a pixel-level accuracy. In the second step, the phase correlation is computed again in the Fourier domain.

$$G(\omega_x,\omega_y)=\left(\frac{F_{tst}(\omega_x,\omega_y)F^*_{src}(\omega_x,\omega_y)}{\left|F_{tst}(\omega_x,\omega_y)F^*_{src}(\omega_x,\omega_y)\right|}\right)$$

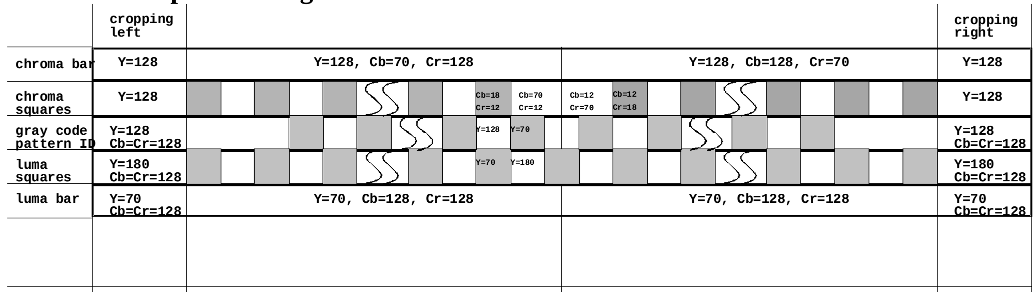
However, this time, it is used in conjunction with a linear regression that fits a line to the phases of the lower frequency components. The calculation results in an estimation of the group delay. Using the Fourier transform property

$$f(x+m, y+n) = \exp(jnu+jmv)F(u,v),$$

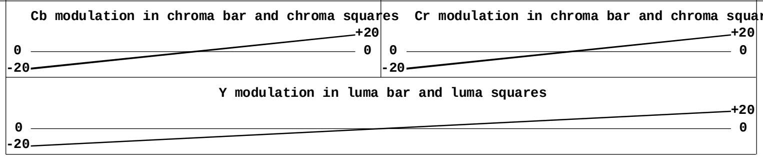
this phase delay is converted to subpixel shift estimates.

## APPENDIX: Structure of Alignment Stripe Pattern

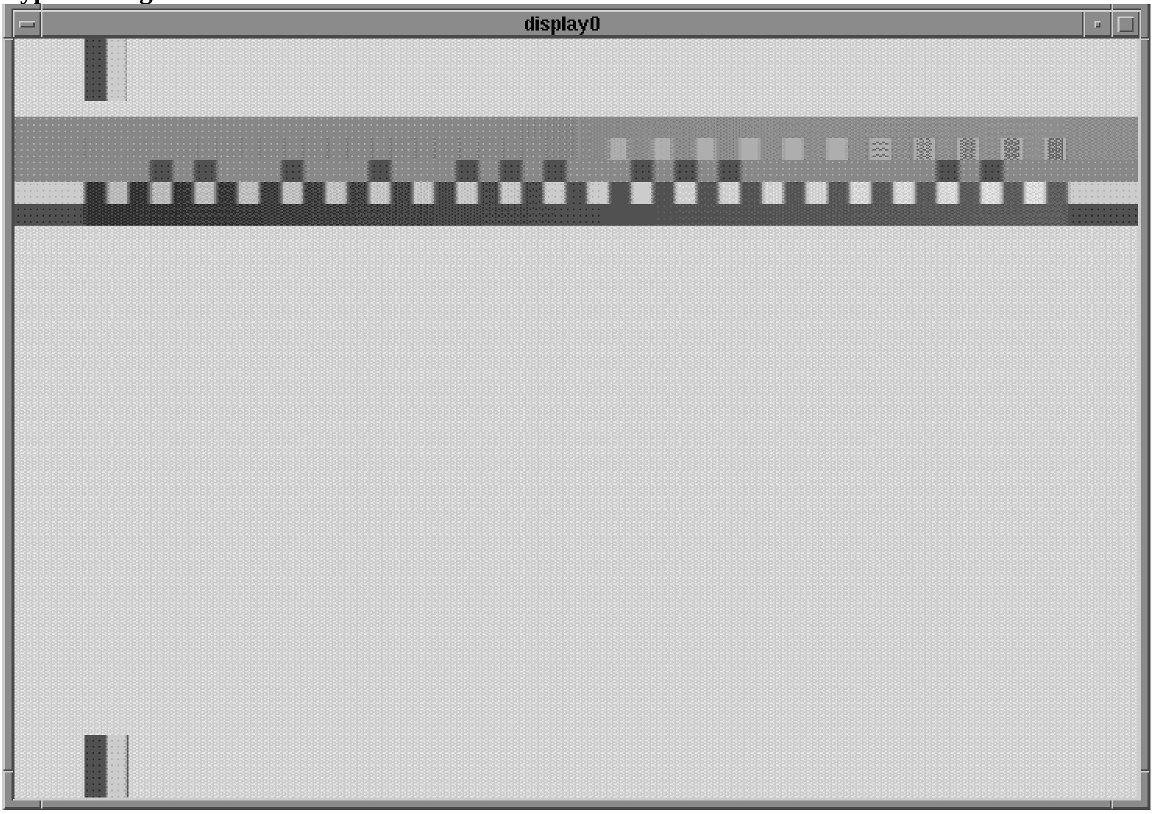
### Detailed description of Alignment bar for Pattern A:



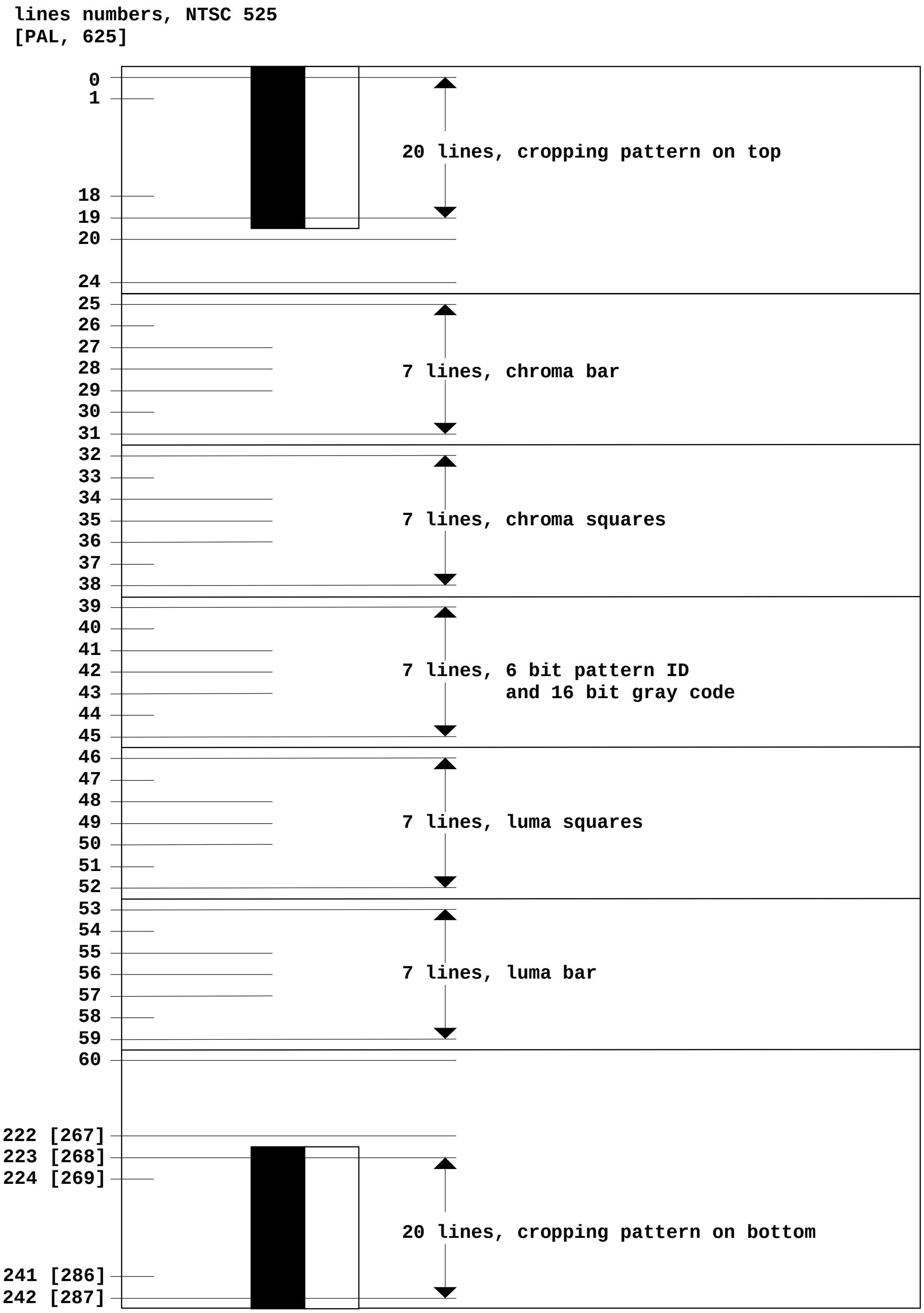
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| pixel address | 0 | 89 | 90 | 719 | 720 | 1349 | 1350 | 1439 |
| pixel count: | 1 | 90 | 91 | 720 | 721 | 1350 | 1351 | 1440 |
| Y<br>number: | 45 | | 315 | | 315 | | 45 | |
| pixel count | 1 | 45 | 46 | 360 | 361 | 675 | 676 | 720 |
| Cb<br>numbers: | 23 | | 157 | | 158 | | 22 | |
| pixel count | 1 | 23 | 24 | 180 | 181 | 338 | 339 | 360 |
| Cr<br>numbers: | 22 | | 158 | | 157 | | 23 | |
| pixel count | 1 | 22 | 23 | 180 | 181 | 337 | 338 | 360 |



### Typical image:



**Further details of test pattern:**



## ANNEX 4: Discussion on the basic test design within VQEG

The range of quality that is to be examined in this test is extremely large. Regarding the issue of how to deal with the subjective testing, two possibilities have been discussed, namely two separate tests and one integrated test.

Two separate tests, one with high bit rate HRCs (called HQ test) and one with low bit rate HRCs (called LQ test) with some overlapping conditions, are intended to avoid compression of subjective results at the end points. This is based on an assumption that rating experiments are sensitive to the range of distortions being presented and that combining all conditions would reduce differences among the high quality conditions. The overlapping conditions are introduced to check for consistency of the two tests. The separate test, however, does not imply a real split into a high quality part and a low quality part. We will have a broad quality range (unimpaired original through the lowest bit rate HRC) in the LQ test and a narrow quality range (unimpaired original through the middle bit rate HRC) in the HQ test. Further, since the picture quality is a product of HRC and sequence, the split based on only the bit rate does not guarantee the split of picture quality.

On the other hand, some members of VQEG presented experimental results which indicate DSCQS's stability. In the study of contextual effect [6], it has been shown that DSCQS is hardly affected by context being presented under a condition that both test series have the same and relatively wide total range of quality. Another study shows that DSCQS is not much influenced by the total range of quality either, at least when expert viewers are used [7]. These data suggest that it is not necessary to separate the test into HQ and LQ tests. From these results some members of VQEG drew the conclusion that a well-balanced mixture of high quality and low quality conditions is an appropriate design in order to minimize contextual effect.

After intensive discussions, it has been adopted that the two separate LQ and HQ tests with some overlapping conditions should be conducted. This design is not perfect but is a compromise to help achieve most of the goals of this test.

[6] ITU-R WP 11E, “Investigation of contextual effects”, 1997

[7] ITU-R WP 11E, “Comparison of reliability of the DSIS (RBU) mtehod and the DSCQS method with limited impairment range”, 1996

## Revision History

In the following, the modifications of the various versions of the subjective test plan are described.

**Version 2:**

- chapter 3.1: notice included that the use of the VQEG test sequences shall be restricted to the VQEG evaluation technical tests
- chapter 3.2: HRC list: interchange of HRCs 15 and 16

**Version 3:**

- Annex 3: updated version of the description of the normalization processing